Predicting Protein Binding Sites

F. Edward Boas
Russ Altman
Biomedical Informatics
Stanford University School of Medicine

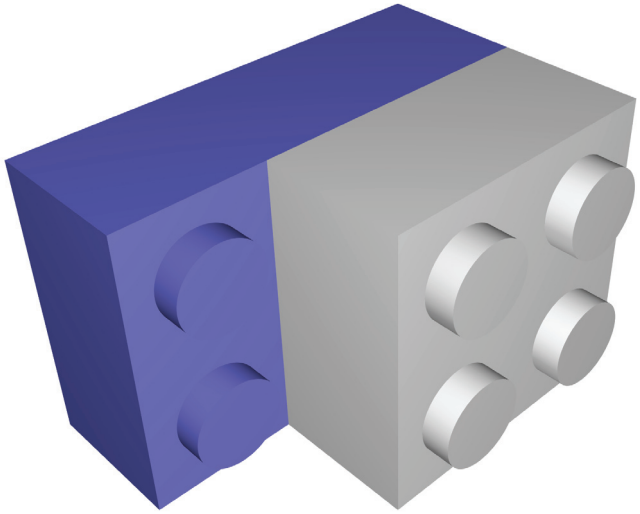
24 May 2000

CELL BIOLOGY DEPENDS ON PROTEIN INTERACTIONS

Protein interactions:

Provide structural support
Assemble subunits of enzymes
Transmit information in cell signaling pathways
Underly cellular / viral / bacterial adhesion

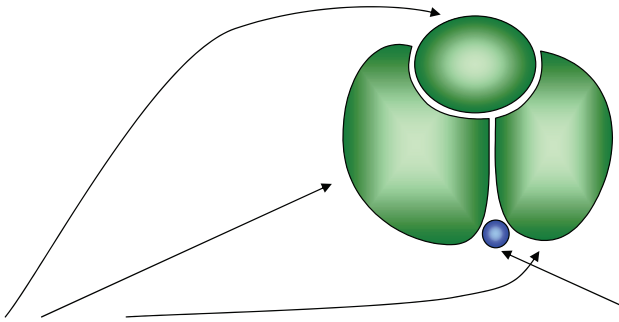
ARE PROTEINS LIKE LEGOS?



How do the physical properties of protein binding interfaces differ from the rest of the protein's surface?

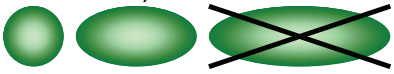
Are these differences sufficient to **predict** binding sites without even knowing the binding partner?

SELECTING CRYSTAL STRUCTURES



Crystal structure must contain at least two proteins† that are:

- 1000 – 2000 atoms (excluding hydrogen)
- > 90% standard amino acids
- globular (largest dimension ≤ 2 × smallest dimension)



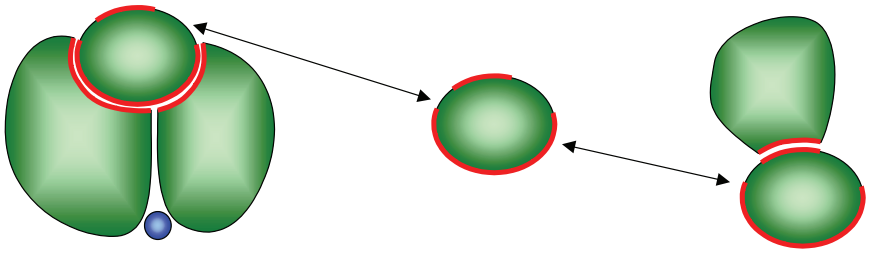
Fewer than 100 atoms (excluding hydrogen) in other chains

These criteria identified 690 crystal structures in the Protein Data Bank.

† "protein" refers to a single polypeptide chain, or multiple covalently connected polypeptide chains

IDENTIFYING ATOMS INVOLVED IN BINDING

Match protein chains in co-crystal structures with individually solved structures

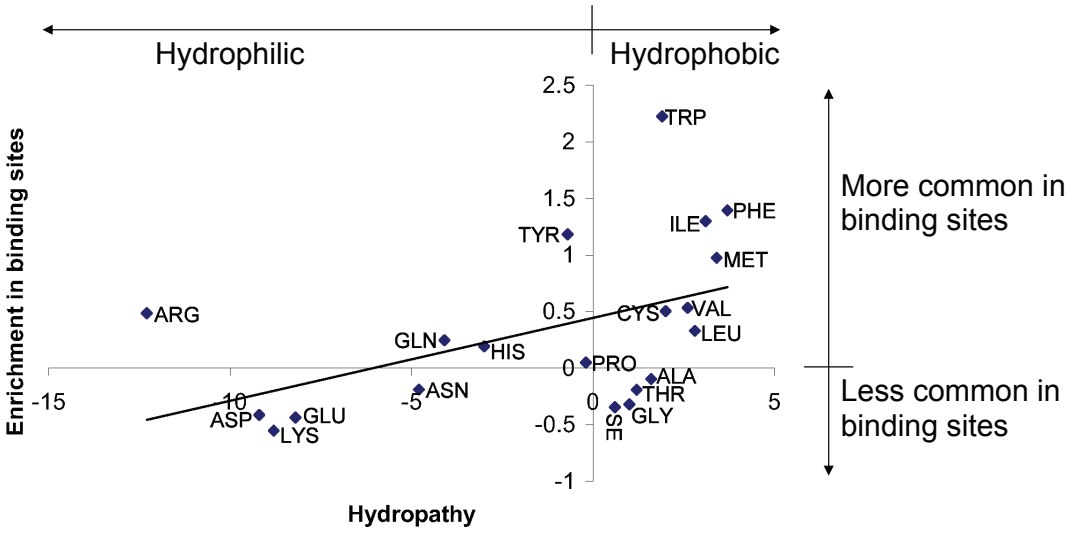


- **Binding sites** are solvent-accessible atoms near another protein chain
- **Non-binding sites** are other solvent-accessible atoms

These criteria identified 4800 distinct protein chains, containing 5.5×10^5 solvent accessible atoms. Of these:

- 1.2×10^5 are binding sites
- 4.3×10^5 are non-binding sites

AMINO ACIDS CONTENT

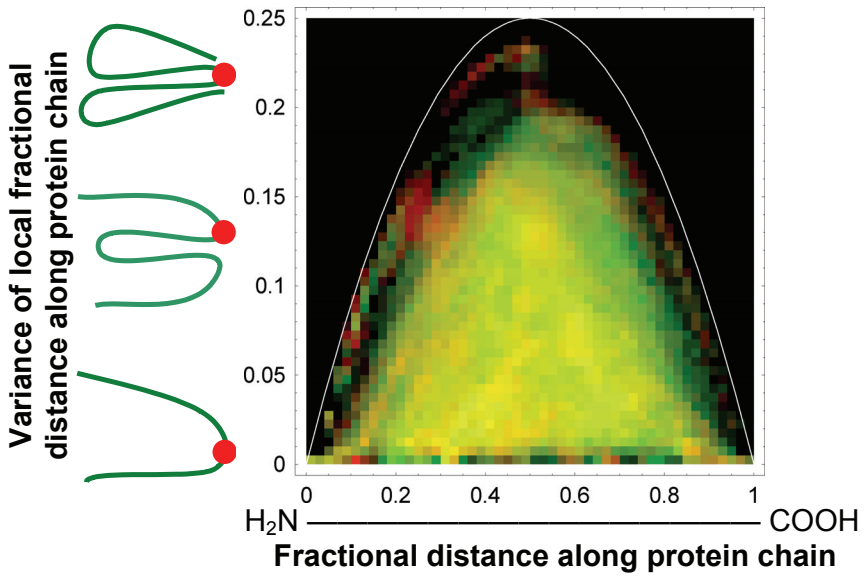


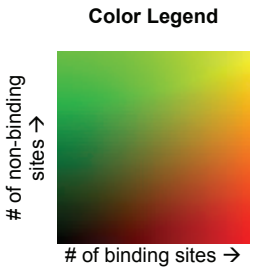
Hydrophobic atoms are more likely to bind another protein, but this is only a weak correlation

"**Hydropathy**" is free energy of amino acid transfer from hydrophobic environment (assumed dielectric constant 2) to water, in kcal/mol

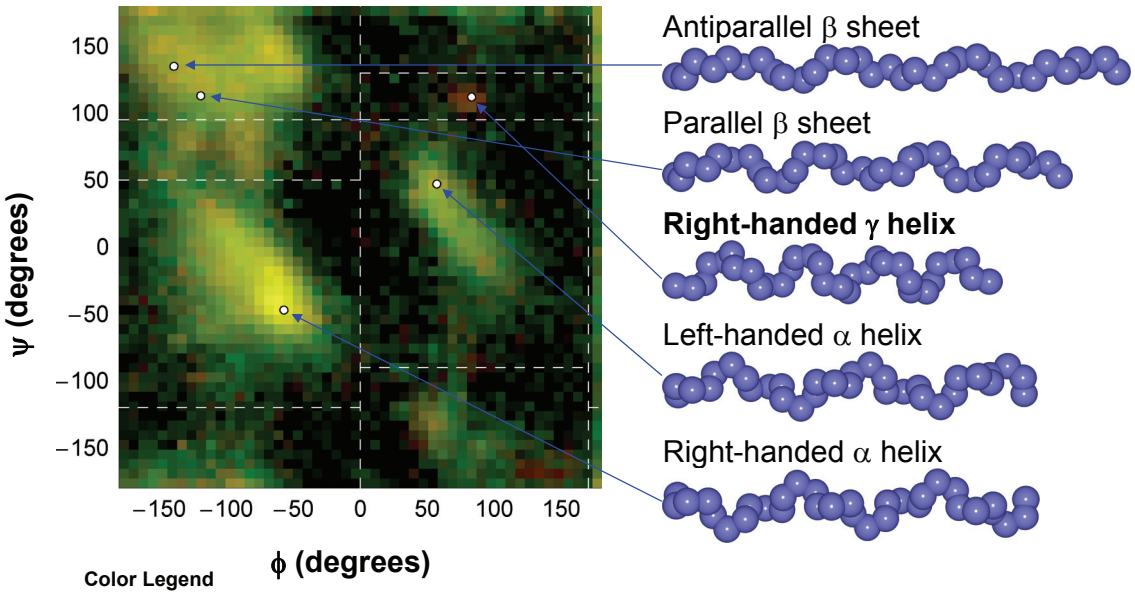
"**Enrichment in binding sites**" is weighted by the number of solvent-accessible atoms

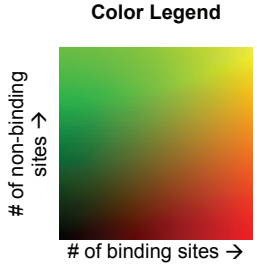
Binding Sites Often Contain Loops from Different Parts of a Peptide Chain





Secondary Structure

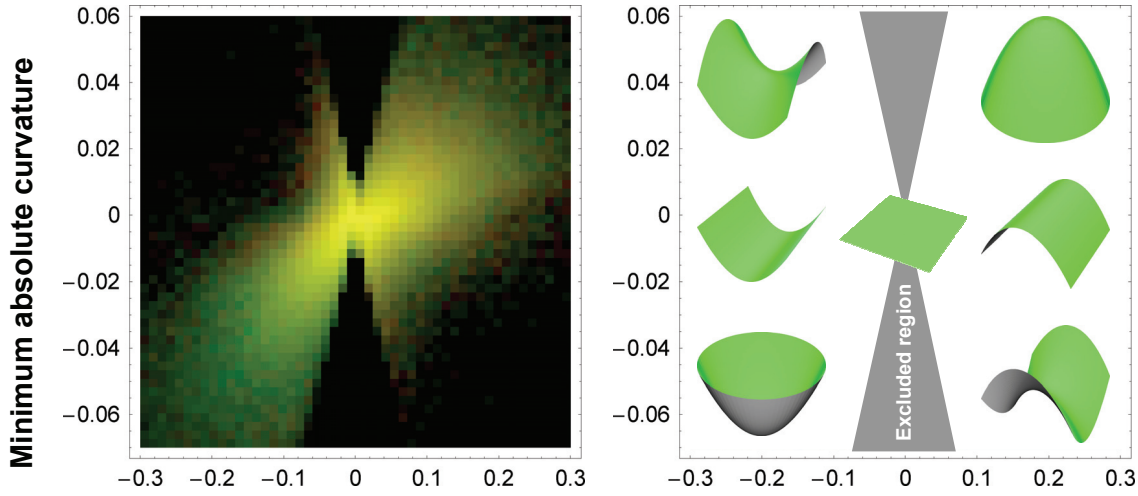




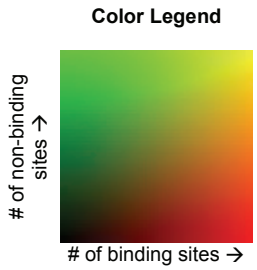
Ramachandran plot shows that secondary structure correlates with binding capability. In particular, the tightly wound, sterically unfavorable **right-handed γ helix** is primarily found in protein binding interfaces

Surface Curvature

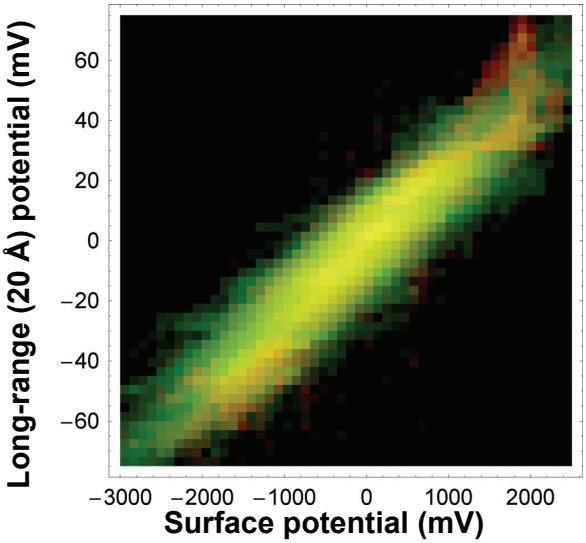
Units of curvature are $Å^{-1}$. Convex surfaces have positive curvature; concave surfaces have negative curvature

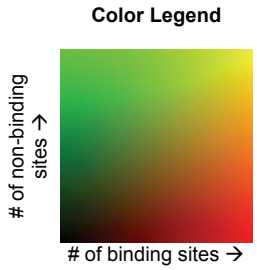


- Saddle points are more likely to be binding sites
- Indentations in the protein surface are less likely to be binding sites



Electrostatic Potential



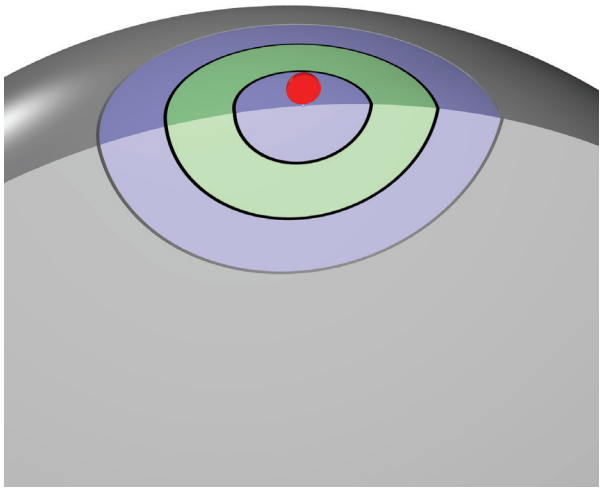


- Large potentials are associated with binding sites
- Correlation is stronger for long-range potential than for surface potential, suggesting that a protein's dipole is favorably oriented with its binding site

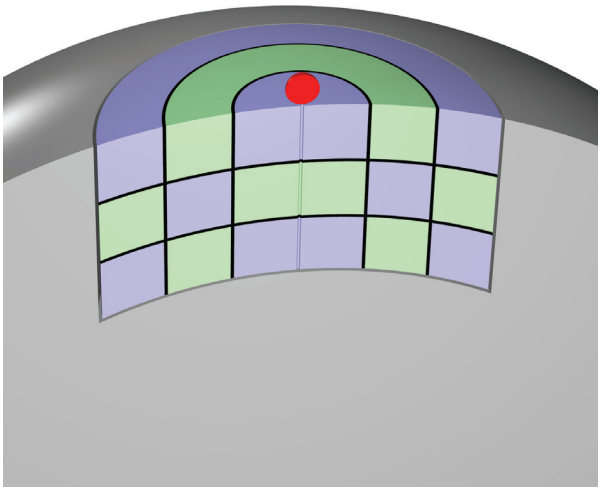
Examining Amino Acid Configurations

To examine the 3D configuration of amino acids around a site of interest, count amino acids in various 3D "bins"

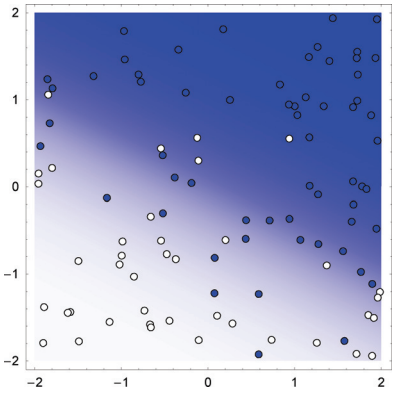
Spherical shells



Cylindrical shells



Logistic Regression



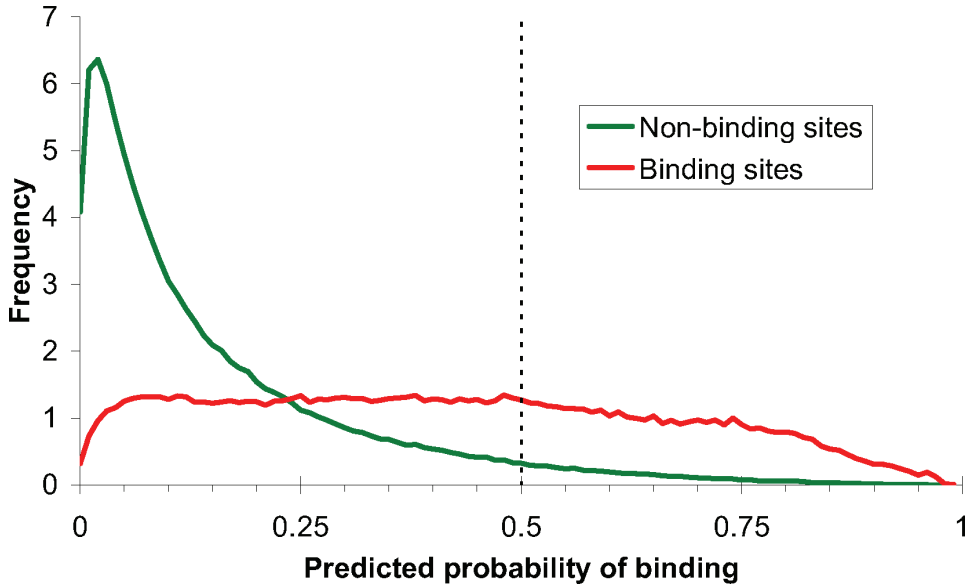
Predict probability of being a blue point as:

$1/(1+e^{c0 + c1\,x1 + c2\,x2})$

Use logistic regression on the following 263 variables to predict binding sites:

- Curvature
- Bumpiness
- Solvent-accessible fraction
- Amino acid
- Amino acid counts in spherical shells
- Amino acid counts in cylindrical shells
- Fractional N→C distance
- Variance local fractional N→C distance
- Surface electrostatics
- Long-range electrostatics
- Secondary structure

Predicting Protein Binding Sites

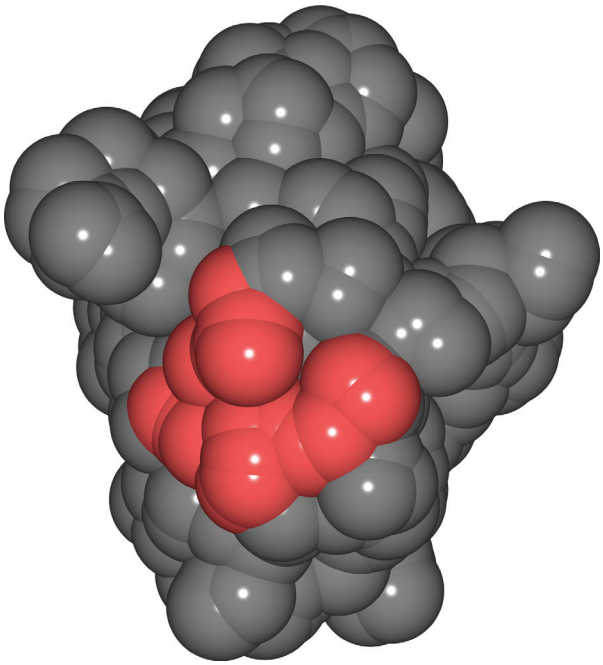


Using a cutoff binding probability of 0.5:

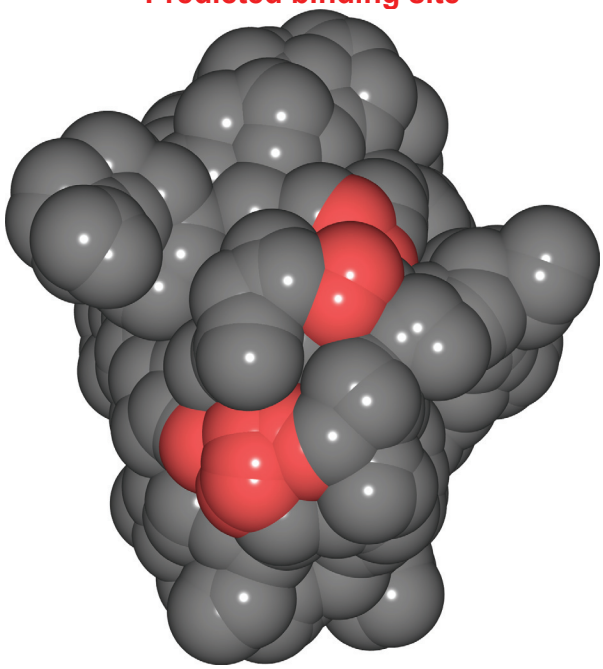
- 15% false negatives
- 34% false positives

Sindbis Virus Capsid Protein

Actual dimer interface



Predicted binding site

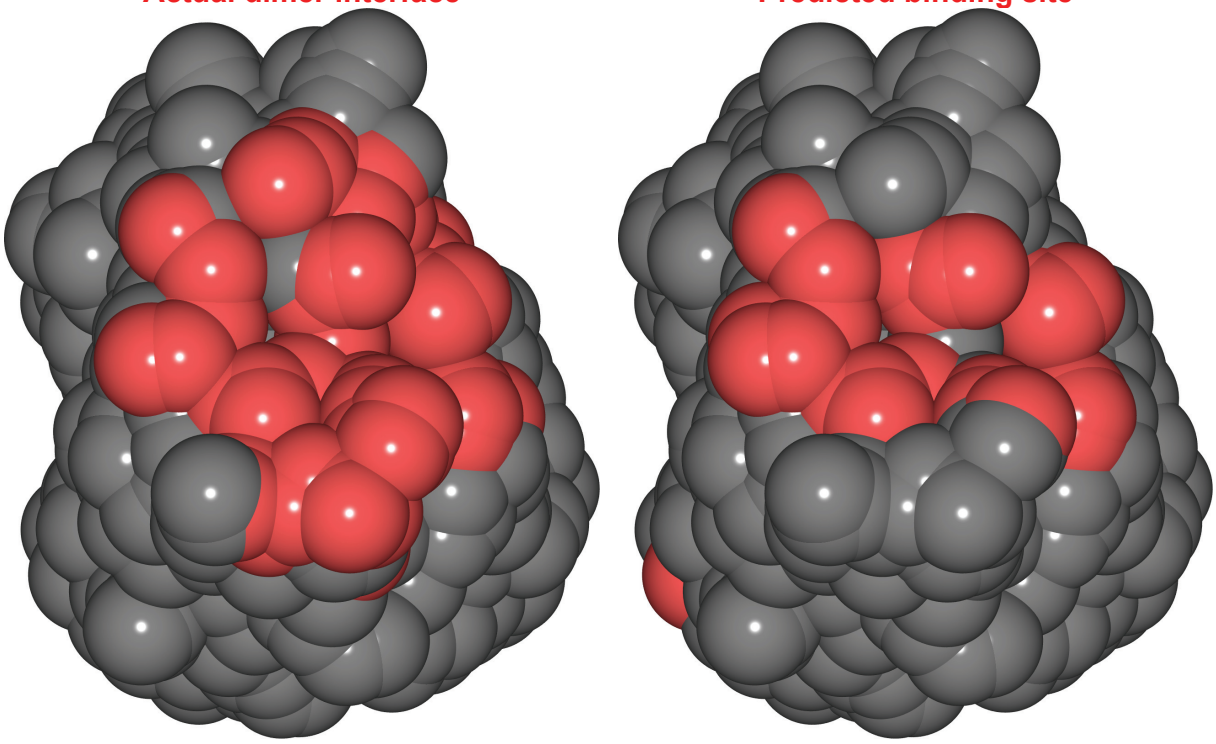


PDB codes: 2SNW, chains A and B; matched with 1KXA

Triose Phosphate Isomerase

Actual dimer interface

Predicted binding site

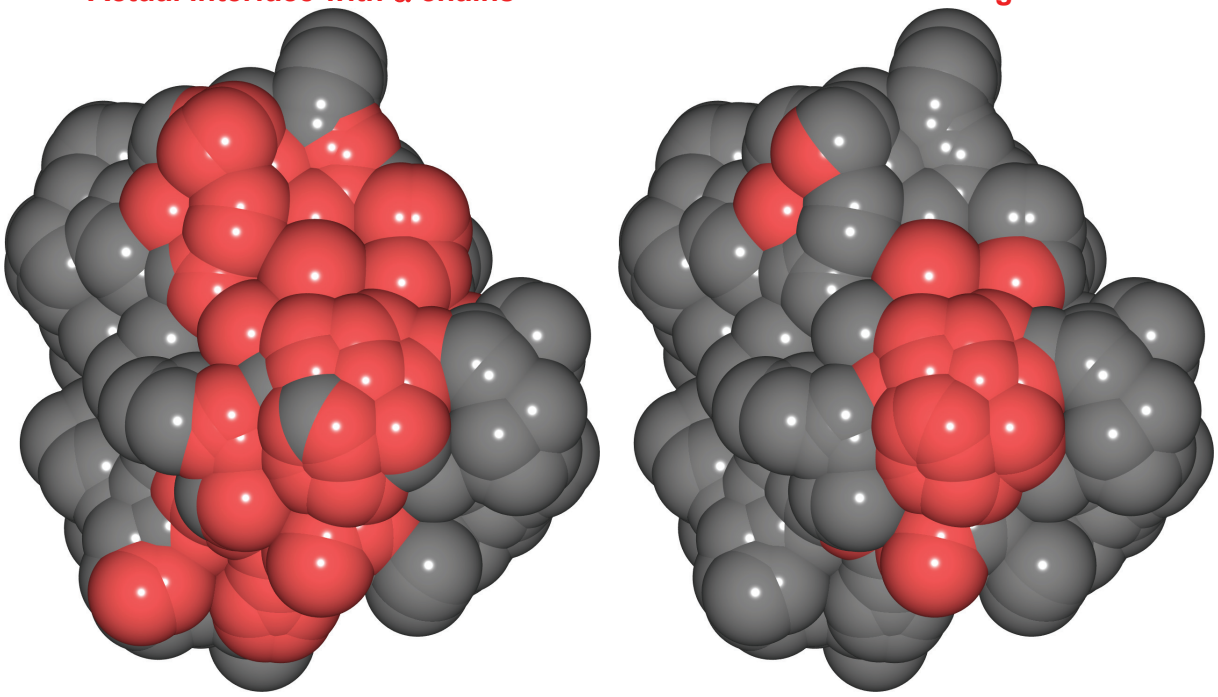


PDB code: 1TPH

Hemoglobin β Chain

Actual interface with α chains

Predicted binding site



PDB code: 4HHB